CCIN2P3

cnrs

**Centre de Calcul**

de l'Institut National de Physique Nucléaire
et de Physique des Particules

# FJPPL – Japan-France workshop on computing technologies

# *Ceph at CC-IN2P3*

**Adrien Georget & Loïc Tortay**

# Ceph @CC: brief time scale

- 2017: cluster for OpenStack @CC (RBD)
- 2019: hosted cluster for IFB (bioinformatics institute, OpenStack)
- 2020: cluster for LSST (CephFS)
- 2021: cluster LSST expansion (+60% capacity)
- 2021: cluster for WoK (RBD+CephFS+S3)
- 2022: cluster for GPFS replacement (CephFS)

# Ceph @CC: clusters

- OpenStack: Nautilus (Ceph v14), 600 TB, 16 disk servers (disks + SSDs & SSDs only), 258 OSDs, replication
- *IFB: Pacific (v16), 500 TB, 4 disk servers (disks + NVMe), 64 OSDs, replication*
- WoK: Quincy (v17), 150 TB, 3 disk servers, 72 OSDs, replication
- LSST: Octopus (v15), 7 PB, 32 disk servers (disks + SSDs), 640 OSDs, erasure coding (8+2) for data & replication for metadata
- SPS: Pacific, 6 PB, 30 disk servers (disks + SSDs), 720 OSDs, erasure coding (8+2), container based deployment (cephadm)
- 3 MONs (physical nodes) on all clusters

# Ceph @CC: hardware configurations

- Dell R730xd, R740xd, R740xd2, R440: 67 servers
- HPE A4200 & DL360: 33 servers
- Dell R740xd2 (LSST cluster, x32):
  - 2x Intel Xeon Silver 4210, 128 GiB
  - 20x 12 TB SAS-NL disks, 4x 1.6 TB SAS SSDs
  - 2x 10G Eth
- HPE A4200 (SPS cluster, x30), similar to LSST Dell R740xd2, except:
  - 16x 14 TB SAS-NL disks, 6x 1 TB SAS SSDs
  - 2x 10/25G Eth

# Ceph @CC: use cases

- RBD for OpenStack
- CephFS:
  - very light use for the OpenStack cluster (Spark, etc.), Manila mostly unused for production (by end users)
  - WoK preproduction
  - somewhat large scale use for LSST & SPS
- RGW very lightly used for both OpenStack & WoK

# Ceph @CC: CephFS/LSST

[1/3]

- 1.1 PiB migrated off GPFS & Isilon to CephFS
- 12 disk servers expansion (+1.8 PiB usable) :
  - 11 days data rebalance (~1.2 PiB)
  - no disruption of users jobs
- Single-thread I/O performance
- Minor bus (documentation, MGR)
- Static mount instead of automont used initially (on computing nodes)
- CephFS (vs GPFS) resilience to end users
- Hardlinks (Anaconda, ...)

# Ceph @CC: CephFS/LSST [2/3]

- (Multi-)MDS stability (initially 2 active + 1 standby) :
  - some clients access blocked for unclear reasons
    ⇒ forcible MDS restart/failover
  - latency spikes seen by users (and monitoring)
  - *scrub* (MDS not OSD)
  - ⇒ since early 2022, single active MDS (2 standbys)
  - `mds_cache_memory_limit` set to 32 GiB
- Scrub rate reduction to cope w/ activity and cluster size
- Grafana & Prometheus monitoring
- *ceph-users@ceph.io* mailing-list

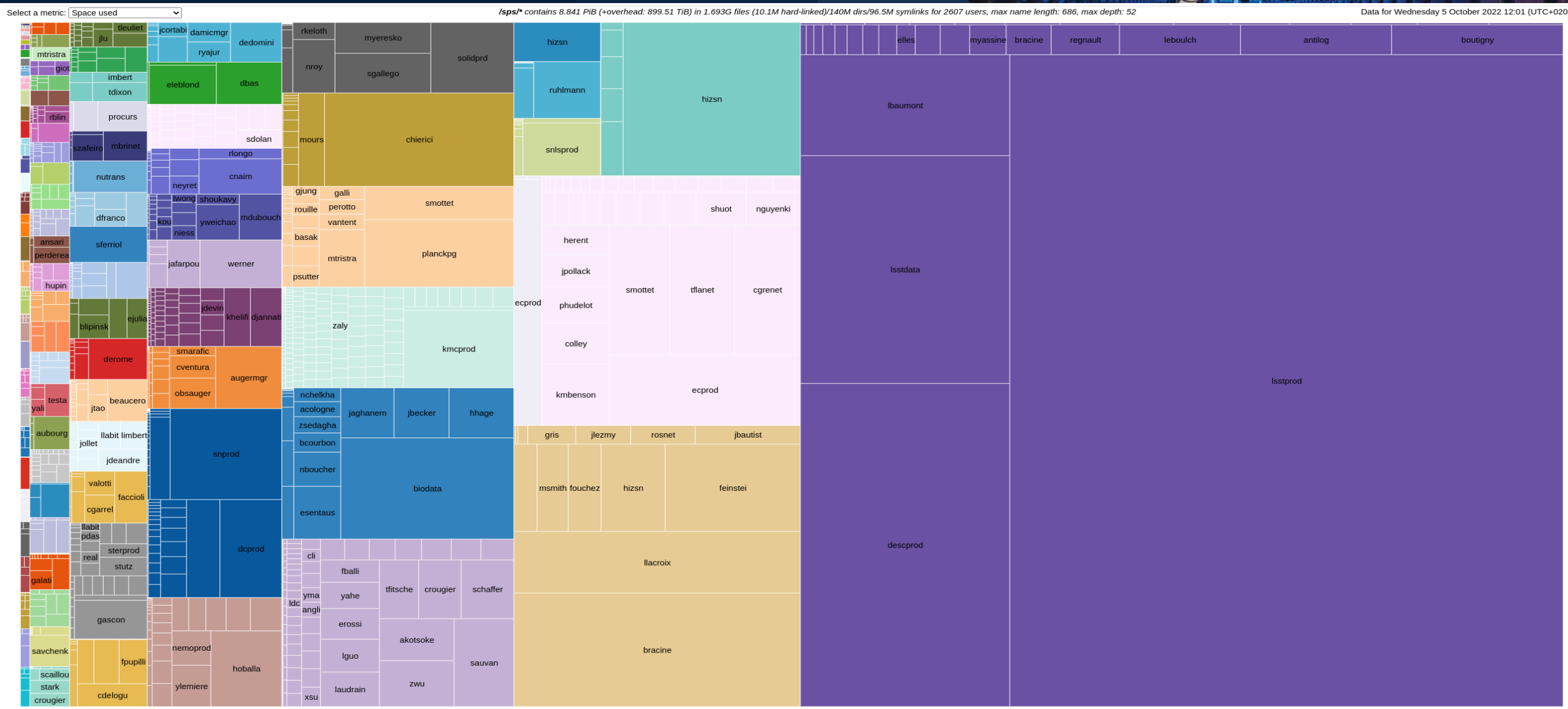
# Ceph @CC: CephFS/LSST [3/3]

- Good resilience to a capricious disk server
  - `upmap-remapped.py` by Dan van der Ster (CERN)
- User (single) directory web export
- Daily metadata collection for accounting & space management

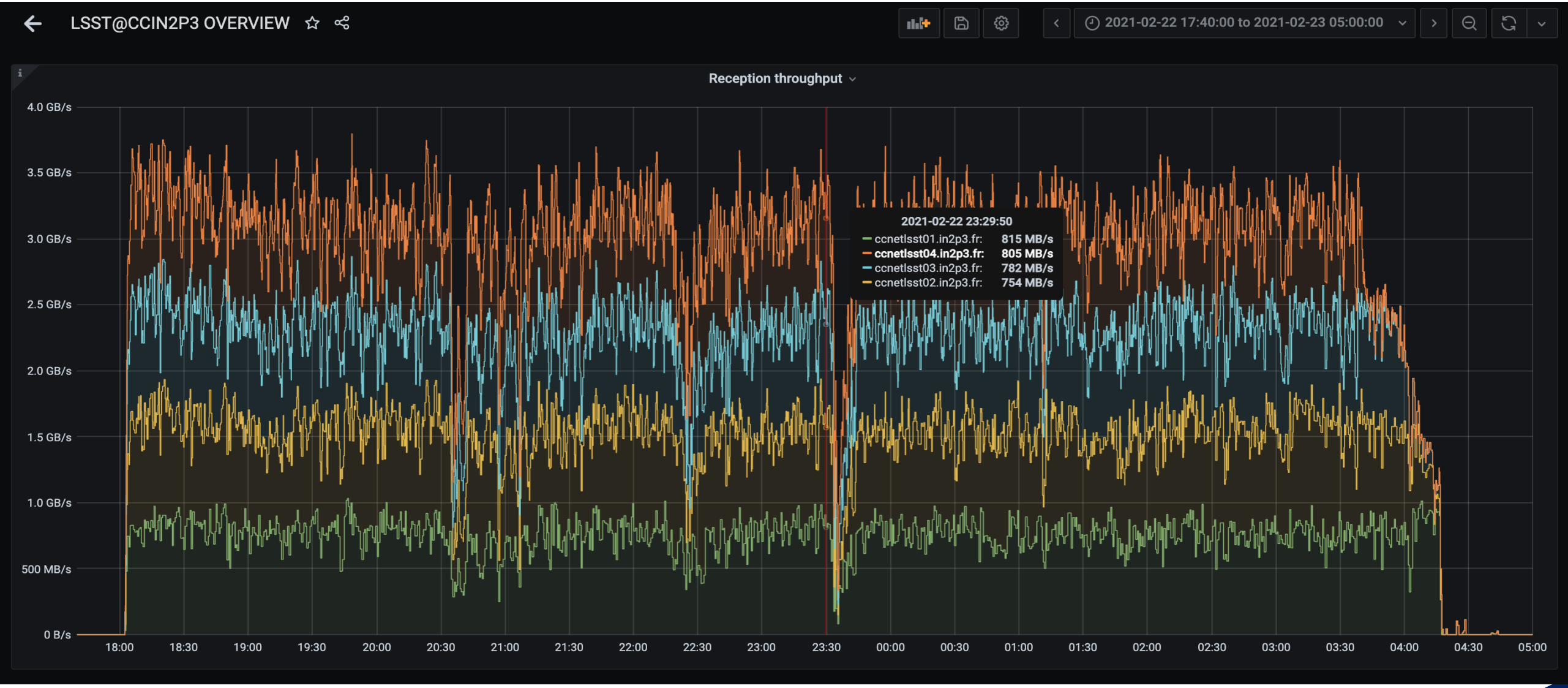
# Ceph @CC: Issues with Ceph Pacific/Quincy & EL7

- Upgrading to Ceph Pacific requires Python 3 (thus EL8)
  ⇒ EL7 clients (i.e. login & computing nodes) left with Ceph Octopus
- New mount syntax in Ceph Quincy effectively prevents access from older clients (using kernel mount)
  ⇒ SPS cluster was downgraded from Quincy to Pacific to allow EL7 clients access

# Ceph @CC: accounting/statistics



# Ceph @CC: LSST transfers CC-IN2P3 ⇔ NERSC



# Ceph @CC: Grafana Ceph

# Ceph @CC : Grafana CephFS

# Ceph @CC : Grafana CephFS

# Ceph @CC: Grafana Ceph activity heatmap

# Ceph @CC: Elastic Search CephFS slow ops